CLINICAL RESEARCH

e-ISSN 1643-3750
© Med Sci Monit, 2026; 32: e953573
DOI: 10.12659/MSM.953573

Received: 2026.03.27
Accepted: 2026.06.16
Available online: 2026.06.23
Published: 2026.XX.XX

Subgroup Differences in Agreement Between an Algorithm Guided Large Language Model and Routine Emergency Department Triage

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

ADEF 1 Ali Halıcı
CD 2 Ezgi Cesur
B 1 Fikret Çelik

1 Department of Emergency Medicine, Faculty of Medicine, Kütahya Health Sciences University, Kütahya, Türkiye
2 Department of Emergency Medicine, Kütahya City Hospital, Kütahya, Türkiye

Corresponding Author: Ali Halıcı, Department of Emergency Medicine, Faculty of Medicine, Kütahya Health Sciences University, Kütahya Health Sciences University, Faculty of Medicine,Evliya Çelebi Yerleşkesi, Tavşanlı Yolu 10. km, Kütahya, Türkiye, Phone: +90 506 641 70 31, e-mail: ali.halici@ksbu.edu.tr

Financial support: None declared
Conflict of interest: None declared

Background: Large language models (LLMs) are increasingly discussed as decision-support tools in emergency care, but their agreement with routine emergency department (ED) triage and subgroup behavior remain insufficiently characterized. We evaluated an algorithm-guided LLM against routine ED triage with emphasis on subgroup heterogeneity and safety-relevant discordance.

Material/Methods: This retrospective study included 1960 adult ED visits with complete triage data. A standardized prompt provided age, sex, chief complaint, comorbidities, systolic/diastolic blood pressure, heart rate, oxygen saturation, temperature, and Glasgow Coma Scale. The LLM assigned 1 triage category within a 5-level Emergency Severity Index–based system (green, yellow-1, yellow-2, red-1, red-2). Outputs were compared with routine ED triage. Performance for urgent vs non-urgent classification was assessed using AUC, sensitivity, specificity, positive predictive value, negative predictive value, F1, and accuracy. Five-level agreement was assessed using quadratic weighted Cohen's kappa and accuracy. Discordance (lower- and higher-acuity LLM vs routine triage) was analyzed across prespecified subgroups.

Results: LLM achieved 71.3% five-level accuracy and substantial agreement with routine triage (weighted $\kappa = 0.824$). For urgent/non-urgent classification, AUC was 0.768, sensitivity 0.630, specificity 0.906. Lower- and higher-acuity discordance rates were 9.8% and 18.9%. Discordance varied across subgroups; lower-acuity assignments vs routine triage were more frequent in older adults, trauma, and diabetes, while infectious presentations showed the highest concordance.

Conclusions: The algorithm-guided LLM showed substantial concordance with routine ED triage but non-uniform subgroup discordance, particularly lower-acuity assignments in patients with older age, diabetes, and trauma. As routine triage served as an operational comparator rather than a gold standard, findings reflect agreement with local practice, not definitive accuracy or safety. Prospective outcome validation is required.

Keywords: Large Language Models • Emergency Medicine • Triage • Artificial Intelligence

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/953573

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Introduction

Emergency departments (EDs) worldwide continue to experience rising patient volumes and increasing overcrowding, placing substantial pressure on already limited healthcare resources [1,2]. In this high-demand environment, triage serves as a critical gatekeeping process, ensuring that patients with time-sensitive and life-threatening conditions receive prompt evaluation and intervention. Accurate triage not only plays a central role in early clinical decision making but also directly influences patient outcomes, resource utilization, and overall ED efficiency, particularly in resource-constrained and overcrowded settings [3]. Errors or delays in triage may lead to missed opportunities for early treatment in critically ill patients or unnecessary allocation of high acuity resources to non-urgent cases, thereby exacerbating overcrowding and workflow inefficiencies.

In recent years, artificial intelligence (AI) and machine learning–based approaches have gained increasing attention as potential tools to support triage decision making in emergency care [4,5]. By leveraging routinely collected triage time data such as vital signs, demographic characteristics, and chief complaint narratives, data-driven triage models aim to provide rapid and consistent risk stratification at the point of presentation [6]. Recent ED machine learning work also demonstrates the feasibility of personalized prediction for downstream resource utilization (eg, electrocardiogram use), supporting the broader role of data driven decision support in emergency care [7]. In particular, natural language processing (NLP) techniques enable the structured interpretation of free-text chief complaints, which represent a rich yet underutilized source of clinical information at triage. Several studies have reported moderate to good performance of AI-assisted triage systems, with discrimination metrics often comparable to traditional triage scales [4,8,9]. These findings suggest that AI has the potential to augment clinical judgment by improving consistency and efficiency, particularly in busy ED settings, where cognitive load and time constraints are substantial.

Despite growing interest in large language model (LLM)-based triage systems, important gaps remain in our understanding of how these systems perform in clinically complex patient groups. Most existing studies focus primarily on overall accuracy or global discrimination metrics, which provide limited insight into performance across vulnerable populations. In particular, the influence of advanced age and chronic disease burden on LLM-based triage decisions have not been sufficiently characterized. Older adults and patients with multiple comorbidities often present with atypical or nonspecific symptoms and may not exhibit the classical physiological responses associated with acute illness, a challenge that is well recognized in geriatric emergency care [10,11]. These limitations are particularly relevant in real-world ED settings, where triage decisions must often be made rapidly with incomplete or nonspecific clinical information. These clinical features raise the possibility that LLM-based models relying mainly on structured inputs such as vital signs and symptom patterns may share similar limitations to conventional triage approaches. As a result, models that demonstrate good average performance may still be less reliable in situations in which patient vulnerability is highest [12]. Furthermore, lower-acuity LLM assignment compared with routine ED triage relative to routine ED triage remains insufficiently explored in much of the current AI triage literature, limiting a comprehensive understanding of discordance in real-world settings.

The incremental contribution of the present study is 3-fold. First, we evaluated an algorithm-guided, prompt-based LLM constrained by a locally implemented 5-level triage framework based on Emergency Severity Index principles, providing contextual validation within a national rule–based triage setting rather than a generic AI classification task. Second, beyond global performance metrics, we focused on lower-acuity LLM assignment compared with routine ED triage and higher-acuity LLM assignment compared with routine ED triage as safety relevant discordance relative to routine ED triage. Third, we examined prespecified subgroup heterogeneity across age, major comorbidities, and presenting complaint categories to identify boundary conditions in which agreement may deteriorate despite acceptable overall performance. Accordingly, the contribution of this study is not limited to overall validation, but also includes boundary condition testing and a safety discordance oriented framing of LLM assisted triage evaluation.

Therefore, the primary objective of this study was not merely to measure overall agreement between an LLM-based triage system and routine ED triage. Rather, we aimed to test whether agreement and safety-relevant discordance varied across prespecified clinically distinct patient subgroups. We hypothesized that the algorithm-guided LLM would show substantial overall agreement with routine ED triage across the 5-level triage scale, and that lower-acuity LLM assignment compared with routine ED triage relative to routine ED triage would be higher in predefined vulnerable subgroups, particularly older adults, patients with diabetes mellitus, and trauma-related presentations. To address this, we examined overall 5-level agreement as well as subgroup-specific patterns of lower-acuity LLM assignment compared with routine ED triage, higher-acuity LLM assignment compared with routine ED triage, and discrimination according to age, major comorbidities, and presenting complaint categories.

Material and Methods

Study Design and Setting

This retrospective observational study was conducted in a tertiary care ED. The study evaluated the performance of

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

an LLM-based triage system using routinely collected triage time data. The study was approved by the Kütahya Health Sciences University Non-Interventional Clinical Research Ethics Committee (decision No. 2025/10-25; Date: 11.08.2025). All data were anonymized prior to analysis, and informed consent was waived due to the retrospective design.

Study Population

Adult patients (≥18 years) presenting to the ED during the study period were eligible for inclusion. From the electronic medical records, we retrieved ED visits with complete documentation of the prespecified core triage variables required for model inference (age, vital signs, and chief complaint). Because these variables were complete for all included records, no visits were excluded for missing data, yielding a final analytic sample of 1960 ED visits.

Data Collection and Variables

Routinely collected triage time data were extracted from electronic medical records. These included the following: demographic characteristics (age and sex); vital signs at presentation (systolic and diastolic blood pressure, heart rate, and pulse oximetry); free-text chief complaint narratives; documented comorbid conditions, including diabetes mellitus, hypertension, cardiovascular disease, cerebrovascular disease, asthma/chronic obstructive pulmonary disease, chronic kidney disease, and cancer; and presenting complaint categories, grouped into predefined clinical domains (eg, cardiovascular, respiratory, neurological, gastrointestinal/abdominal, trauma, infectious, psychiatric, musculoskeletal, dermatologic. and gynecological/urological).

Age was categorized into 3 predefined groups (18-44 years, 45-64 years and ≥65 years) for stratified analyses.

LLM-Based Triage Inference Protocol

The triage system was implemented as an algorithm-guided, prompt-based LLM using ChatGPT 5.2 (OpenAI), accessed through the standard ChatGPT web interface. All LLM inferences were performed between August 1, 2025, and October 1, 2025. The model was used as an off-the-shelf system and was not fine-tuned, locally trained, or adapted using the study dataset.

Prior to case-level predictions, the model was provided with a locally implemented 5-level ED triage algorithm based on Emergency Severity Index principles, including definitions and decision rules for the 5 acuity categories. The local ED triage protocol used in our institution is aligned with this locally implemented 5-level framework.

For each ED visit, the model received a standardized structured input in a fixed order: age, sex, chief complaint, comorbidities, systolic blood pressure, diastolic blood pressure, heart rate, oxygen saturation, body temperature, and Glasgow Coma Scale. An example input was: "21, female, headache, none, SBP 116, DBP 81, HR 90, SpO_2 98, T 36.1, Glasgow Coma Scale 15". The model was instructed to return exactly 1 triage label from the predefined set: green, yellow-1, yellow-2, red-1, or red-2. No explanatory text, reasoning, confidence score, or alternative label was requested. No routine ED triage label or ED triage decision was included in the model input.

Prompt stability was ensured by using the same system instruction, the same 5-level triage framework, the same fixed input order, and the same output restriction for all cases. Each ED visit was evaluated independently, and no conversation history was carried over between cases. No iterative prompting, feedback, chain of thought elicitation, clinician adjudication during inference, or post hoc selection among multiple outputs was performed.

Each case was evaluated in a single-pass format. Outputs were not regenerated or repeated for performance optimization. The only prespecified exception was that if an output did not exactly match one of the predefined triage labels, the identical prompt would be re-issued once, solely to obtain a valid label format. In the present dataset, all outputs matched the predefined label set, and no cases required exclusion for invalid output.

The standard ChatGPT web interface did not allow manual control or reporting of deterministic sampling parameters such as temperature, top-p, seed, or related decoding settings. Therefore, the platform's default user interface settings were used consistently for all queries. Because commercial LLM systems may undergo version updates and may show output variability over time, the reported agreement metrics should be interpreted as specific to the model version, interface, inference period, and fixed prompt protocol used in this study.

All model inputs were de-identified and contained no direct patient identifiers. Free-text chief complaints were included as recorded at triage, but no additional clinical notes, diagnostic results, radiology reports, treatment information, disposition outcomes, or follow-up data were provided to the model.

For LLM inference, the prompt included only the chief complaint text as documented at triage, together with the prespecified structured variables; no additional narrative notes or complete symptom lists were entered. When multiple complaints were documented in the triage complaint field, the predominant triage concern recorded by ED staff was retained as the chief complaint input. For subgroup categorization of presenting

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

complaints, predefined clinical domains were applied using structured review of the chief complaint field; when overlapping features were present, trauma was prioritized.

Reference Standard and Triage Classification

The reference standard was the routine triage acuity assigned by experienced ED triage personnel according to the local triage protocol. This operational triage decision was used as the comparator because it reflects real-world workflow in the study setting. For the primary binary analysis, ED triage categories were dichotomized as urgent (red-1/red-2) vs non-urgent (green/yellow-1/yellow-2). Lower-acuity LLM assignment compared with routine ED triage was defined as the LLM assigning non-urgent when ED triage was urgent; higher-acuity LLM assignment compared with routine ED triage was defined as the LLM assigning urgent when ED triage was non-urgent. Routine ED triage was used as a pragmatic operational comparator reflecting real-world workflow in the study setting; however, it was not treated as an outcome-validated gold standard for true clinical urgency. Accordingly, all reported performance measures should be interpreted as agreement or discordance relative to routine ED triage rather than definitive clinical correctness.

This 5-level framework is the routine triage system used in our setting and is conceptually comparable to other 5-level triage systems used internationally, although specific criteria and acuity thresholds are not identical.

Repeat Run Stability Analysis

Repeat-run (test-retest) stability was evaluated in a prespecified subset of 500 visits by re-running the identical prompt template and input schema. Agreement between the initial and repeat LLM outputs was assessed using quadratic weighted Cohen's kappa for the 5-level triage labels and Cohen's kappa with absolute percent agreement for the binary urgent/non-urgent classification.

Outcomes

The primary prespecified analysis was a binary urgent/non-urgent comparison based on routine ED triage (urgent: red-1/red-2; non-urgent: green/yellow-1/yellow-2). Binary performance was evaluated using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), F1 score, and accuracy. A complementary ordinal analysis evaluated agreement across all 5 triage categories using quadratic weighted Cohen's kappa, 5-level accuracy, and a 5 × 5 confusion matrix.

Planned summary outputs also included the overall proportion of correctly classified cases across the 5-level triage scale and the overall lower-acuity LLM assignment compared with routine ED triage and higher-acuity LLM assignment compared with routine ED triage proportions in the full cohort. These overall results are presented in the main Results text and in the corresponding figures and tables.

Prespecified hypothesis-driven subgroup analyses were performed to examine heterogeneity in agreement and safety relevant discordance across clinically distinct patient groups. These subgroup analyses focused on age strata, major comorbidities, and presenting complaint categories, with particular emphasis on whether lower-acuity LLM assignment compared with routine ED triage relative to routine ED triage was higher in predefined vulnerable subgroups.

Statistical Analysis

Continuous variables were summarized as mean ± standard deviation, and categorical variables as counts and percentages. Normality of continuous variables was assessed using the Shapiro-Wilk test. Normally distributed variables were compared using the independent-samples *t* test, whereas non-normally distributed variables were compared using the Mann-Whitney U test. Analyses were structured into 3 prespecified components: (1) overall agreement across the 5-level triage scale, (2) binary urgent/non-urgent classification performance, and (3) exploratory subgroup analyses according to age, comorbidities, and presenting complaint categories. Five-level agreement between the LLM and routine ED triage was evaluated using quadratic weighted Cohen's kappa, 5-level accuracy, and a 5 × 5 confusion matrix.

For binary urgent/non-urgent analyses, urgent ED triage (red-1/red-2) was treated as the positive class and non-urgent ED triage (green/yellow-1/yellow-2) as the negative class. Diagnostic performance was assessed using sensitivity, specificity, PPV, NPV, F1 score, and overall accuracy. ROC analysis and AUC with corresponding 95% CIs were calculated for the overall cohort and age strata. For comorbidity and presenting complaint subgroups, analyses focused on descriptive concordance and lower-acuity and higher-acuity discordance patterns relative to routine ED triage. For ROC and AUC analyses, the 5 triage categories were encoded as an ordinal score (green = 1, yellow-1 = 2, yellow-2 = 3, red-1 = 4, red-2 = 5), with higher scores indicating greater urgency. For binary point-estimate metrics, LLM outputs of red-1 or red-2 were classified as predicted urgent, whereas outputs of green, yellow-1, or yellow-2 were classified as predicted non-urgent.

Lower-acuity LLM assignment compared with routine ED triage was defined as assignment of a non-urgent category when routine ED triage was urgent, and higher-acuity LLM assignment compared with routine ED triage as assignment of an urgent category when routine ED triage was non-urgent.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Subgroup analyses were prespecified but exploratory and descriptive. They were performed to examine potential heterogeneity in agreement and discordance patterns relative to routine ED triage, rather than to establish statistically confirmed between-group differences or effect modification. No formal interaction testing was performed. Therefore, subgroup findings were interpreted as hypothesis-generating descriptive patterns and were not used to make definitive inferential claims regarding subgroup-specific model performance. Group comparisons were conducted using the chi-square test or Fisher's exact test for categorical variables, and the *t* test or Mann-Whitney U test for continuous variables, as appropriate. A 2-sided *P* value < 0.05 was considered statistically significant.

To assess across-run output stability, a repeat-run (test-retest) analysis was performed in a prespecified subset of 500 visits by re-running the identical prompt template and input schema. Agreement between runs was quantified using quadratic weighted Cohen's kappa for the 5-level triage labels and Cohen's kappa with percent agreement for the binary urgent/non-urgent classification.

All analyses were performed using IBM SPSS Statistics and Python.

Results

Cohort Characteristics

A total of 1960 ED visits were included. All model outputs matched the predefined label set; therefore, no cases were excluded due to invalid responses. Urgent cases (red-1/red-2) accounted for 431 of 1960 visits (22.0%), while non-urgent cases accounted for 78.0% of the cohort. Baseline characteristics by ED triage urgency are summarized in **Table 1**.

Overall 5-Level Agreement

Five-level agreement between the LLM and routine ED triage is presented in **Figure 1** as a 5 × 5 confusion matrix. Quadratic weighted Cohen's kappa indicated strong ordinal agreement (κ = 0.824, 95% CI: 0.807-0.840). Most disagreements occurred between adjacent acuity levels (eg, yellow-1 vs yellow-2), whereas discordance spanning non-urgent to urgent categories accounted for the lower-acuity LLM assignment compared with routine ED triage and higher-acuity LLM assignment compared with routine ED triage patterns described below.

Binary Urgent/Non-Urgent Classification Performance

ROC analysis was performed to evaluate the ability of the LLM-based triage system to distinguish urgent (red-1/red-2) from non-urgent (green/yellow-1/yellow-2) cases. In the overall cohort, the LLM system achieved an AUC of 0.768, sensitivity of 0.630, specificity of 0.906, PPV of 0.672, NPV of 0.888, F1 score of 0.650, and overall accuracy of 0.845. Age-stratified performance metrics are summarized in **Table 2**.

Exploratory Subgroup Analyses

Prespecified exploratory subgroup analyses were performed according to age, major comorbidities, and presenting complaint categories to describe potential heterogeneity in agreement and discordance patterns relative to routine ED triage. These analyses were descriptive, and no formal interaction testing was performed. Therefore, the subgroup findings should be interpreted as hypothesis-generating patterns rather than statistically confirmed between-group differences. To further characterize the clinical contexts in which lower-acuity LLM assignment compared with routine ED triage and higher-acuity LLM assignment compared with routine ED triage occurred, we examined stratified subgroup results. In prespecified age-stratified descriptive analyses, AUC was numerically lower in older age groups, with the lowest estimate observed in patients aged 65 years and older. Among patients aged 18 to 44 years, the AUC was 0.774 (95% CI: 0.723-0.823), compared with 0.760 (95% CI: 0.715-0.801) in those aged 45 to 64 years and 0.705 (95% CI: 0.662-0.744) in those aged 65 years and older. Descriptively, lower-acuity LLM assignment compared with routine ED triage was more frequent in the oldest age group, whereas concordant LLM–routine triage assignment was most frequent in younger adults.

In the overall study population, the LLM system correctly classified 71.3% of cases, while 9.8% were lower-acuity LLM assignment compared with routine ED triaged, and 18.9% were higher-acuity LLM assignment compared with routine ED triaged. Age-stratified lower-acuity LLM assignment compared with routine ED triage and higher-acuity LLM assignment compared with routine rates of ED triage are summarized in **Figure 2**.

When stratified by age, patients aged 18 to 44 years demonstrated the highest proportion of concordant LLM routine triage assignments, accompanied by relatively low rates of lower-acuity LLM assignment compared with routine ED triage (5.4%) and higher-acuity LLM assignment compared with routine ED triage (17.9%). In the 45 to 64 years age group, lower-acuity LLM assignment compared with routine ED triage increased to 10.3%, while concordant assignment decreased to 66.5%. Among patients aged 65 years and older, lower-acuity LLM assignment compared with routine ED triage was more frequent (18.1%), whereas concordant assignment was comparatively lower than in younger adults (65.6%).

Higher-acuity LLM assignment compared with routine ED triage and lower-acuity LLM assignment compared with routine

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Table 1. Baseline characteristics by emergency department triage urgency (urgent vs non-urgent).

	Non-urgent	Urgent	*P* value
Sex, n			< 0.001
Female	841 (82.5%)	178 (17.5%)	
Male	688 (70%)	253 (30%)	
Age (years)			< 0.001
18-44	876 (90.9%)	88 (9.1%)	
45-64	368 (72.9%)	137 (27.1%)	
≥65	285 (58%)	206 (42%)	
Vital parameters			
Systolic blood pressure (mm Hg)	124 ± 20	126 ± 30	0.990
Diastolic blood pressure (mm Hg)	76 ± 11	75 ± 16	0.010
Heart rate (beats/min)	93 ± 17	93 ± 23	0.035
Pulse oximetry (%)	97 ± 2	94 ± 6	< 0.001
Medical history			
Diabetes mellitus	164 (56.6%)	126 (43.3%)	< 0.001
Hypertension	236 (57.4%)	175 (42.6%)	< 0.001
Cardiovascular disease	128 (43.5%)	166 (56.5%)	< 0.001
Cerebrovascular disease	15 (42.9%)	20 (57.1%)	< 0.001
Asthma/COPD	32 (44.4%)	40 (55.6%)	< 0.001
Chronic kidney disease	9 (40.9%)	13 (59.1%)	< 0.001
Cancer	30 (76.9%)	9 (23.1%)	0.868
Complaint categories			< 0.001
Cardiovascular	128 (42.5%)	173 (57.5%)	
Respiratory	41 (40.6%)	60 (59.4%)	
Neurological	179 (78.9%)	48 (21.2%)	
Gastrointestinal/abdominal	345 (88.9%)	43 (11.1%)	
Trauma	201 (77.6%)	58 (22.4%)	
Psychiatric	24 (66.7%)	12 (33.3%)	
Infectious	367 (94.8%)	20 (5.2%)	
Dermatologic	33 (84.6%)	6 (15.4%)	
Musculoskeletal	146 (96.1%)	6 (3.9%)	
Gynecological/urological	65 (92.9%)	5 (7.1%)	

* Categorical variables were compared using the chi-square test or Fisher's exact test, as appropriate. Continuous variables were compared using the independent-samples *t* test or Mann-Whitney U test, according to data distribution.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

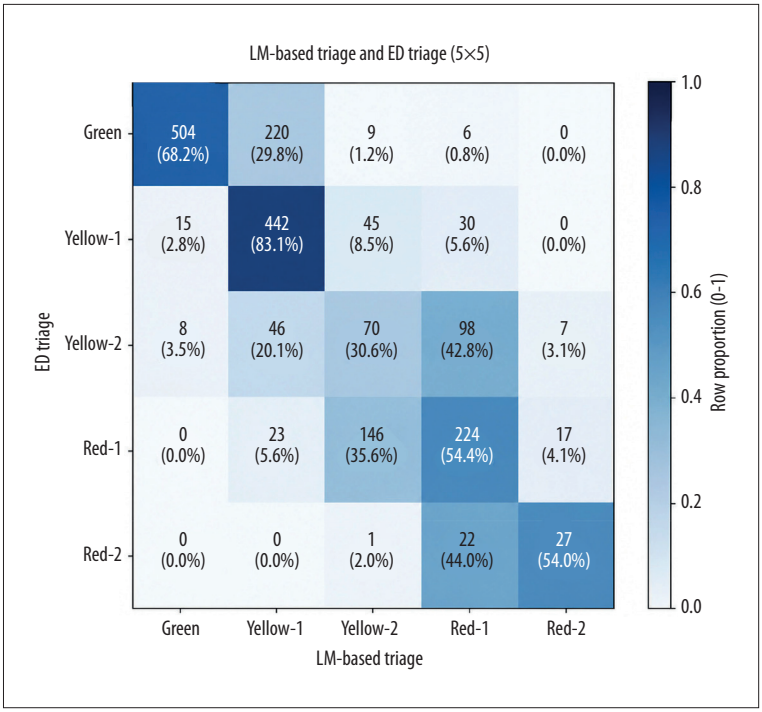

Figure 1. Five-level confusion matrix (5 × 5) comparing routine emergency department (ED) triage (reference) and large language model (LLM)-based triage (predicted) across 5 categories (green, yellow-1, yellow-2, red-1, red-2).

Table 2. ROC performance of large language model (LLM)-based triage for urgent cases by age group.

	AUC (95% CI)	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score
All patients	0.768 (0.745-0.791)	0.845	0.630	0.906	0.672	0.888	0.650
18-44 years	0.774 (0.723-0.823)	0.927	0.587	0.961	0.614	0.957	0.600
45-64 years	0.760 (0.715-0.801)	0.809	0.652	0.868	0.642	0.872	0.647
≥65 years	0.705 (0.662-0.744)	0.716	0.635	0.775	0.718	0.702	0.674

Abbreviations: AUC, area under the receiver operating characteristic (ROC) curve; PPV, positive predictive value; NPV, negative predictive value.

ED triage rates of the LLM-based triage system according to comorbid conditions and presenting complaint categories are shown in **Figure 3**.

In prespecified descriptive subgroup analyses by comorbidity, patients with diabetes mellitus showed a lower proportion of concordant LLM routine triage assignments and a higher frequency of lower-acuity LLM assignment compared with routine ED triage than the overall cohort. By presenting complaint, infectious presentations showed the highest proportion of concordant assignments and comparatively low discordance in the lower-acuity direction, whereas trauma-related presentations showed more frequent lower-acuity LLM assignment compared with routine ED triage.

Among patients with diabetes mellitus, concordant LLM routine triage assignment was observed in 56.6% of cases, with lower-acuity LLM assignment compared with routine ED triage in 19.3% and higher-acuity LLM assignment compared with routine ED triage in 24.1%. In patients with hypertension, concordant assignment was observed in 66.9%, whereas lower-acuity and higher-acuity LLM assignments compared with routine ED triage occurred in 14.6% and 18.5%, respectively. For patients with cardiovascular disease, concordant assignment was observed in 67.3% of cases, with lower-acuity and higher-acuity LLM assignments compared with routine ED triage in 13.9% and 18.7%, respectively.

When stratified by presenting complaint, gastrointestinal presentations showed a lower proportion of concordant LLM

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

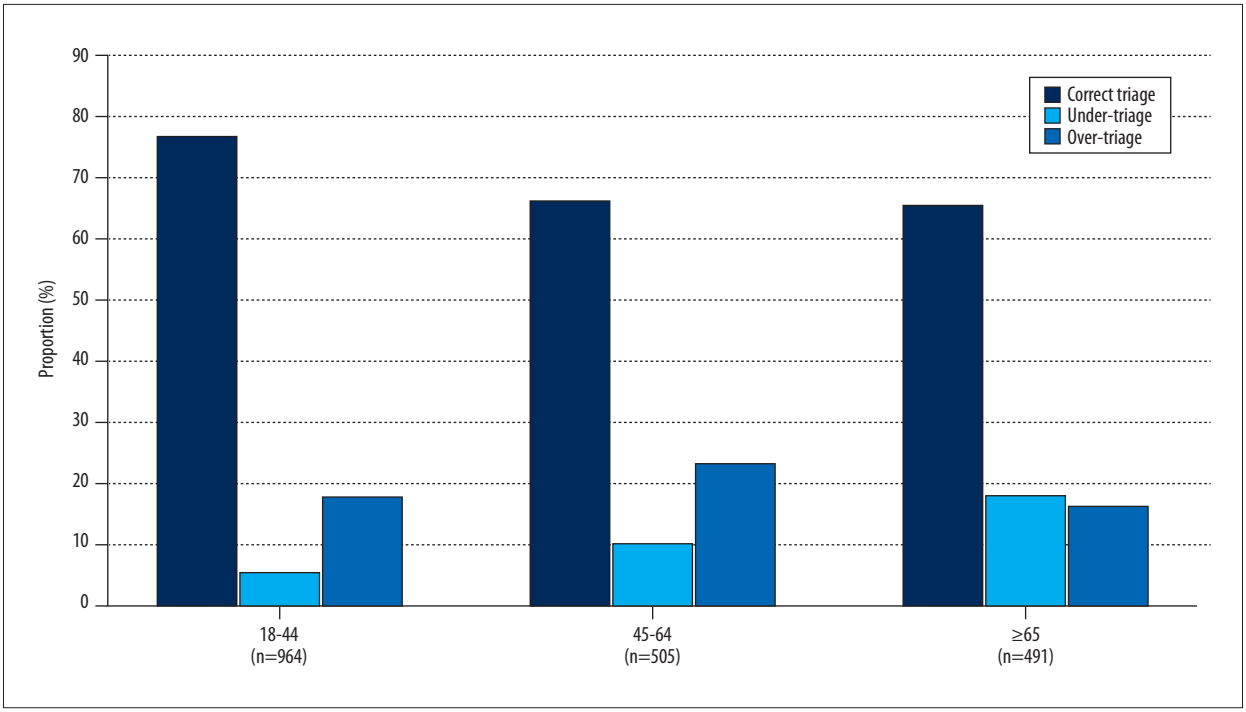

Figure 2. Age stratified higher-acuity LLM assignment compared with routine ED triage and lower-acuity LLM assignment compared with routine ED triage rates of the LLM-based triage system * Subgroup results are descriptive and exploratory. ** Age-stratified concordant LLM–routine triage assignment, lower-acuity LLM assignment compared with routine ED triage, and higher-acuity LLM assignment compared with routine ED triage rates. Error bars indicate Wilson 95% confidence intervals. Numbers indicate subgroup sample size.

routine triage assignments (60.8%), accompanied by more frequent lower-acuity and higher-acuity LLM assignments compared with routine ED triage (14.2% and 25.0%, respectively). Infectious presentations demonstrated the highest proportion of concordant assignments (80.1%), with comparatively low lower-acuity discordance relative to routine ED triage (4.1%) and moderate higher-acuity discordance (15.8%).

Cardiovascular presentations were predominantly concordant with routine ED triage (76.7%), with lower-acuity and higher-acuity LLM assignments compared with routine ED triage in 6.0% and 17.3% of cases, respectively. In contrast, trauma-related presentations showed more frequent lower-acuity LLM assignment compared with routine ED triage (22.0%), despite a concordant assignment rate of 68.0% and a lower frequency of higher-acuity discordance (10.0%) compared with other categories. In neurological presentations, concordant assignment was lower (63.9%), with contributions from both lower-acuity and higher-acuity LLM assignments compared with routine ED triage (10.6% and 25.6%, respectively).

Repeat-Run Stability Analysis

Repeat-run stability was evaluated in a prespecified subset of 500 visits using the identical prompt template and input schema. Across repeated runs, 5-level agreement was excellent, with a quadratic weighted Cohen's kappa of 0.976. For the binary urgent/non-urgent classification, absolute agreement between runs was 96.8%, with a Cohen's kappa of 0.925. These findings indicate high across-run output stability under the fixed inference protocol used in this study.

Discussion

In this retrospective agreement-based evaluation, an algorithm-guided LLM showed substantial concordance with routine ED triage, with high 5-level ordinal agreement (weighted $\kappa = 0.824$), moderate discrimination for urgent vs non-urgent routine triage categories (AUC 0.768), and high specificity relative to routine non-urgent triage assignments (0.906). However, because routine ED triage was used as an operational comparator rather than an outcome-validated gold standard, these findings should be interpreted as agreement with local triage practice rather than evidence of true agreement with routine ED triage or validated safety. ED presentations are inherently heterogeneous, encompassing different age groups, comorbidity profiles, symptom patterns, and levels of acuity. In this context, our findings suggest that LLM-based triage concordance is unlikely to be uniform across all patient groups and

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

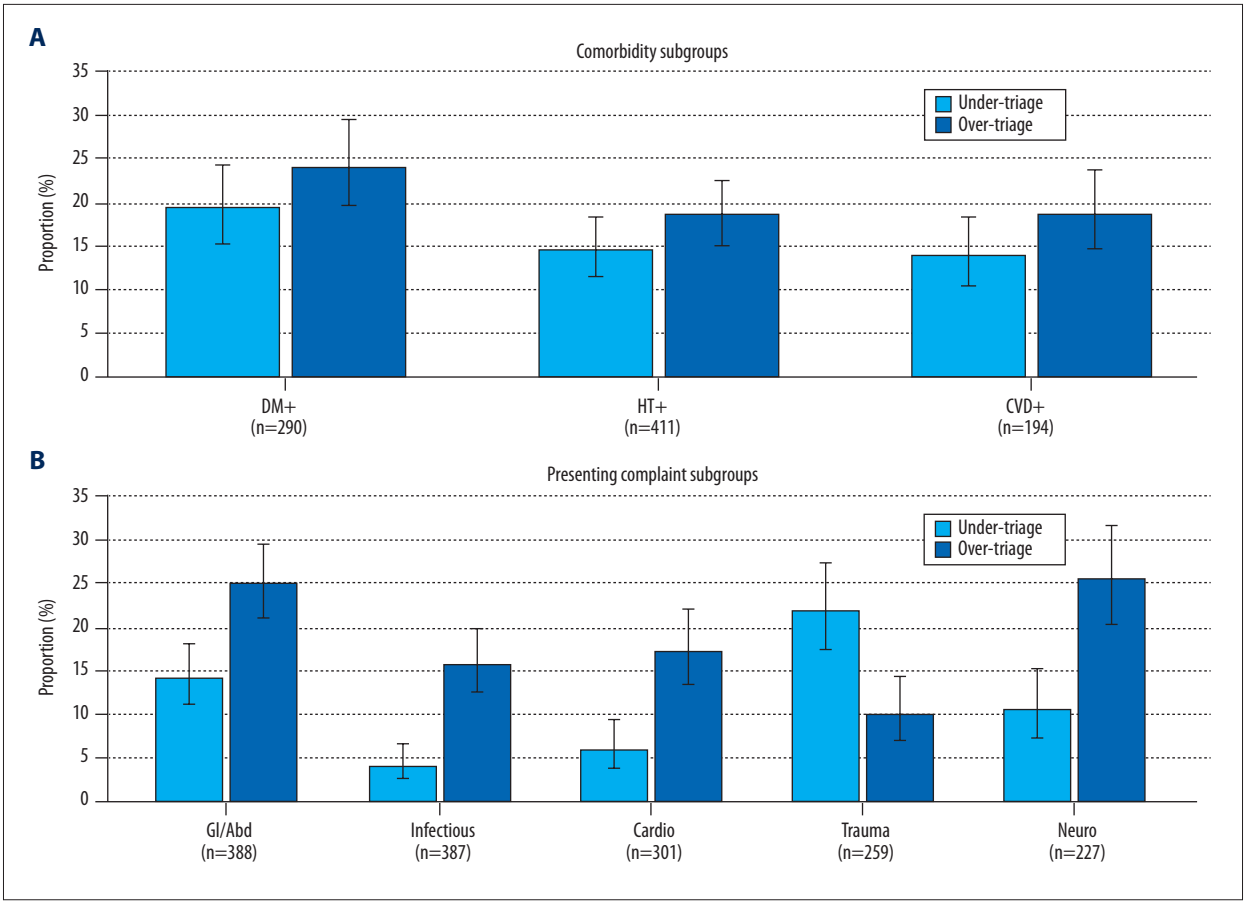

Figure 3. Over and lower-acuity large language model (LLM) assignment compared with routine emergency department (ED) triage by comorbidities and complaint categories. Notes: Values are descriptive subgroup estimates. Higher-acuity LLM assignment compared with routine ED triage and lower-acuity LLM assignment compared with routine ED triage rates by comorbidities and presenting complaint categories. Error bars indicate Wilson 95% confidence intervals. Numbers indicate subgroup sample size. Results are exploratory. Abbreviations: DM, diabetes mellitus; HT, hypertension; CVD, cardiovascular disease; GI/Abd, gastrointestinal/abdominal complaints; Cardio, cardiovascular complaints, Neuro: neurological complaints.

should not be judged solely by overall summary metrics. The higher frequency of lower-acuity LLM assignments compared with routine ED triage among older adults, patients with diabetes mellitus, and trauma-related presentations supports the need for subgroup-sensitive evaluation and outcome-based validation of future LLM-supported triage systems.

In a real-world cohort largely composed of non-urgent routine triage presentations, the LLM showed an overall 5-level concordance of 71.3% and high specificity for non-urgent routine ED triage categories. This pattern suggests that the model frequently aligned with lower-acuity routine triage assignments, which may be relevant in crowded EDs, where unnecessary prioritization can affect resource use and patient flow. However, these findings should not be interpreted as evidence that the model accurately identified true low-acuity clinical states, because no outcome-based reference standard was available. Overall discrimination relative to routine ED triage in our study was broadly consistent with previous triage systems in emergency care [4,5,13-15]. At the same time, binary urgent vs non-urgent metrics alone may not fully capture disagreements occurring within intermediate triage categories. For this reason, 5-level agreement measures such as weighted kappa and the full confusion matrix provide important complementary information when assessing concordance with routine triage practice.

An important consideration in triage evaluation is the balance between lower-acuity LLM assignment compared with routine ED triage and higher-acuity LLM assignment compared with routine ED triage. In most emergency settings, lower-acuity LLM assignment compared with routine ED triage is generally regarded as the more consequential error because urgent patients may experience delayed assessment or treatment, whereas higher-acuity LLM assignment compared with routine ED triage primarily affects resource use and patient flow.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

Many established triage systems therefore accept a degree of higher-acuity LLM assignment compared with routine ED triage in order to reduce missed high-acuity cases [16]. In the present study, the LLM showed an higher-acuity LLM assignment compared with routine ED triage rate of 18.9% and an lower-acuity LLM assignment compared with routine ED triage rate of 9.8%, suggesting a classification profile that favored sensitivity to urgent illness while maintaining reasonable control of unnecessary escalation. Nevertheless, the 5-level confusion matrix showed that some ED red-1/red-2 cases were assigned to intermediate yellow categories, indicating that favorable overall agreement can still coexist with clinically relevant discordance at the high-acuity boundary. These findings highlight the importance of examining misclassification patterns, not only summary accuracy measures, when assessing triage systems.

While ROC analysis provides a useful summary of discrimination relative to routine ED triage, it offers limited insight into the direction and potential implications of discordance. Therefore, we interpreted lower-acuity LLM assignments compared with urgent routine ED triage as the more safety-relevant direction of disagreement. Importantly, this interpretation does not establish true clinical harm or confirm missed high-acuity illness, because outcome data were not used as the reference standard. Rather, it identifies the subgroup contexts in which the LLM diverged from routine ED triage in a direction that would generally be considered more concerning in triage practice.

In descriptive age-stratified analyses, agreement and discrimination relative to routine ED triage were numerically lower in older adults, with the lowest AUC observed in patients aged 65 years and older and a higher frequency of lower-acuity LLM assignment compared with routine ED triage (18.1%). This pattern is clinically plausible, as older adults often present with atypical symptoms, nonspecific complaints, or less pronounced physiological responses during acute illness [17-19]. Similar challenges may also affect LLM-based triage systems that rely primarily on structured vital signs and chief complaint text. In addition, multimorbidity and polypharmacy can further complicate early risk assessment in this population. Taken together, our findings suggest that models performing adequately in the general ED population may be less reliable in older patients unless age-related vulnerability is more explicitly incorporated. Future approaches may benefit from age-adapted calibration strategies, including greater weighting of comorbidity burden or frailty-related indicators.

Our descriptive stratified analyses also suggested lower concordance with routine ED triage in patients with greater comorbidity burden, particularly those with diabetes mellitus, in whom concordant LLM routine triage assignment was 56.6% and lower-acuity LLM assignment compared with routine ED triage was 19.3%. Similar, although less pronounced, patterns were observed in patients with hypertension and cardiovascular disease. These findings are consistent with previous reports showing that medically complex patients are more likely to be lower-acuity LLM assignment compared with routine ED triaged because acute deterioration may be masked by chronic symptoms or atypical presentations [17,20-23]. Triage systems based mainly on vital signs and brief complaint descriptions may therefore underestimate risk in these groups. The higher lower-acuity LLM assignment compared with routine ED triage observed in patients with diabetes mellitus may be especially relevant, as autonomic dysfunction, silent ischemia, and nonspecific presentations are well recognized in this population. Future LLM-based triage models may benefit from more explicit incorporation of comorbidity-related risk factors.

To further explore sources of misclassification, we examined performance across presenting complaint categories. A descriptive contrast was observed between infectious and trauma-related presentations. Infectious complaints showed the highest concordant LLM routine triage assignment rate (80.1%) and one of the lowest lower-acuity LLM assignment compared with routine ED triage rates (4.1%), whereas trauma-related presentations had a substantially higher lower-acuity LLM assignment compared with routine ED triage rate (22.0%). This difference is clinically plausible. Infectious illnesses often produce measurable physiological abnormalities, such as fever, tachycardia, tachypnea, hypoxemia, or hypotension, which are more readily captured by routine triage variables and structured inputs [24-26]. Similar findings have been reported in prior AI-based prediction studies using vital signs and early clinical data [27,28]. In contrast, trauma triage frequently depends on information that may not be fully represented in standard triage datasets, including mechanism of injury, anatomical injury pattern, pain behavior, and visual findings, such as deformity or external bleeding [29]. As a result, LLM-based systems relying mainly on vital signs and complaint text may underestimate severity when overt physiological derangement is initially absent. These findings suggest that trauma-focused adaptations, including structured injury descriptors or integration with rule-based trauma protocols, may improve future triage performance.

Taken together, these descriptive subgroup findings suggest that lower-acuity LLM assignment compared with routine ED triage may be more frequent in certain vulnerable patient groups; however, these observations should be interpreted as hypothesis-generating and require prospective outcome-based validation.

Study Strengths

This study has several important strengths. It was conducted in a large real-world ED cohort reflecting routine triage practice,

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

which increases the practical relevance of the findings. The inclusion of a broad and heterogeneous patient population, encompassing different age groups, comorbidity profiles, and presenting complaint categories, allowed detailed subgroup analyses beyond conventional aggregate performance measures, such as overall accuracy or AUC. As a result, the study was able to identify clinically meaningful variation in performance that may not be apparent from population-level summaries alone.

Another strength is the focus on misclassification patterns rather than overall discrimination alone. In addition to reporting standard agreement metrics, we specifically examined lower-acuity LLM assignment compared with routine ED triage and higher-acuity LLM assignment compared with routine ED triage across clinically relevant subgroups. This approach highlighted vulnerable groups, particularly older adults, patients with diabetes mellitus, and trauma-related presentations, in whom reduced performance may have practical implications. By identifying both the areas of strong agreement and contexts in which further refinement may be needed, the study provides useful direction for the future development and evaluation of more reliable, subgroup-sensitive LLM-based triage systems.

Conclusions

In this retrospective study, an algorithm-guided LLM showed substantial agreement with routine local ED triage; however, agreement and discordance patterns varied across clinically distinct patient subgroups. These findings reflect concordance with routine triage decisions rather than validated clinical urgency, diagnostic accuracy, or patient outcomes. Routine ED triage should therefore be interpreted as an operational comparator, not as an outcome-based gold standard. Overall, the results suggest that subgroup-sensitive evaluation may be as important as aggregate concordance when assessing future LLM-supported triage tools. Prospective outcome-based studies and multicenter external validation are required before clinical implementation can be recommended.

Department and Institution Where Work Was Done

Department of Emergency Medicine, Faculty of Medicine, Kütahya Health Sciences University, Kütahya, Türkiye.

Patient Permission/Consent Declaration

The study was approved by the Kütahya Health Sciences University Non-Interventional Clinical Research Ethics Committee (decision No. 2025/10-25; date: 11.08.2025). The requirement for informed consent was waived due to the retrospective study design and the use of anonymized data.

The study was conducted in accordance with the Declaration of Helsinki.

Availability of Data and Materials

The datasets generated and/or analyzed during the current study are not publicly available due to ethical and privacy restrictions (including clinical free-text fields) and applicable data protection regulations. De identified data required to reproduce the main analyses may be made available from the corresponding author upon reasonable request, subject to institutional approval and a data use agreement.

Declaration of Figures' Authenticity

All figures submitted have been created by the authors who confirm that the images are original with no duplication and have not been previously published in whole or in part.

References:

1. Morley C, Unwin M, Peterson GM, et al. Emergency department crowding: A systematic review of causes, consequences and solutions. PLoS One. 2018;13(8):e0203316
2. Sartini M, Carbone A, Demartini A, et al. Overcrowding in emergency department: Causes, consequences, and solutions – A narrative review. Healthcare (Basel). 2022;10(9):1625
3. Mistry B, Stewart De Ramirez S, Kelen G, et al. Accuracy and reliability of emergency department triage using the emergency severity ındex: An ınternational multicenter assessment. Ann Emerg Med. 2018;71(5):581-87.e3
4. Raita Y, Goto T, Faridi MK, et al. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care. 2019;23(1):64
5. Porto BM. Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review. BMC Emerg Med. 2024;24(1):219
6. Zhang X, Kim J, Patzer RE, et al. Prediction of emergency department hospital admission based on natural language processing and neural networks. Methods Inf Med. 2017;56(5):377-89
7. Wang H, Zhang X. Machine learning for personalized prediction of electrocardiogram (EKG) use in emergency care. J Pers Med. 2025;15(8):358
8. Tschoellitsch T, Seidl P, Böck C, et al. Using emergency department triage for machine learning-based admission and mortality prediction. Eur J Emerg Med. 2023;30(6):408-16
9. Yi N, Baik D, Baek G. The effects of applying artificial intelligence to triage in the emergency department: A systematic review of prospective studies. J Nurs Scholarsh. 2025;57(1):105-18
10. Hong W, Earnest A, Sultana P, et al. How accurate are vital signs in predicting clinical outcomes in critically ill emergency department patients. Eur J Emerg Med. 2013;20(1):27-32
11. Hogervorst VM, Buurman BM, De Jonghe A, et al. Emergency department management of older people living with frailty: A guide for emergency practitioners. Emerg Med J. 2021;38(9):724-29
12. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544-47

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]

13. Karamanlıoğlu A, Demirel B, Tural O, et al. Privacy-preserving clinical decision support for emergency triage using LLMs: System architecture and real-world evaluation. Applied Sciences. 2025;15(15):8412
14. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS One. 2018;13(7):e0201016
15. Colakca C, Ergın M, Ozensoy HS, ety al. Emergency department triaging using ChatGPT based on emergency severity index principles: A cross-sectional study. Sci Rep. 2024;14(1):22106
16. Davis JW, Dirks RC, Sue LP, Kaups KL. Attempting to validate the overtriage/undertriage matrix at a Level I trauma center. J Trauma Acute Care Surg. 2017;83(6):1173-78
17. Alshibani A, Alharbi M, Conroy S. Under-triage of older trauma patients in prehospital care: A systematic review. Eur Geriatr Med. 2021;12(5):903-19
18. Lim JY, Jee Y, Choi SG, et al. Redefining trauma triage for elderly adults: Development of age-specific guidelines for ımproved patient outcomes based on a machine-learning algorithm. Medicina (Kaunas). 2025;61(5):784
19. Ingielewicz A, Szarafińska M, Zając M, et al. Triage and hospitalization outcomes in the geriatric population of an emergency department: A retrospective cohort study comparing the manchester triage system and the emergency severity index. PLoS One. 2025;20(9):e0332304
20. Samaras N, Chevalley T, Samaras D, Gold G. Older patients in the emergency department: A review. Ann Emerg Med. 2010;56(3):261-69
21. Chang DC, Bass RR, Cornwell EE, Mackenzie EJ. Undertriage of elderly trauma patients to state-designated trauma centers. Arch Surg. 2008;143(8):776-82
22. Egodage T, Ho VP, Bongiovanni T, et al. Geriatric trauma triage: optimizing systems for older adults – A publication of the American Association for the Surgery of Trauma Geriatric Trauma Committee. Trauma Surg Acute Care Open. 2024;9(1):e001395
23. Canto JG, Shlipak MG, Rogers WJ, et al. Prevalence, clinical characteristics, and mortality among patients with myocardial infarction presenting without chest pain. JAMA. 2000;283(24):3223-29
24. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-74
25. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med. 2020;46(3):383-400
26. Keep JW, Messmer AS, Sladden R, et al. National early warning score at Emergency Department triage may allow earlier identification of patients with severe sepsis and septic shock: A retrospective observational study. Emerg Med J. 2016;33(1):37-41
27. Goh KH, Wang L, Yeow AYK, et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nat Commun. 2021;12(1):711
28. Țocu G, Lisă EL, Tutunaru D, et al. The potential of artificial ıntelligence in the diagnosis and prognosis of sepsis: A narrative review. Diagnostics (Basel). 2025;15(17):2169
29. Eastridge BJ, Salinas J, McManus JG, et al. Hypotension begins at 110 mm Hg: Redefining "hypotension" with data. J Trauma. 2007;63(2):291-99

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS]